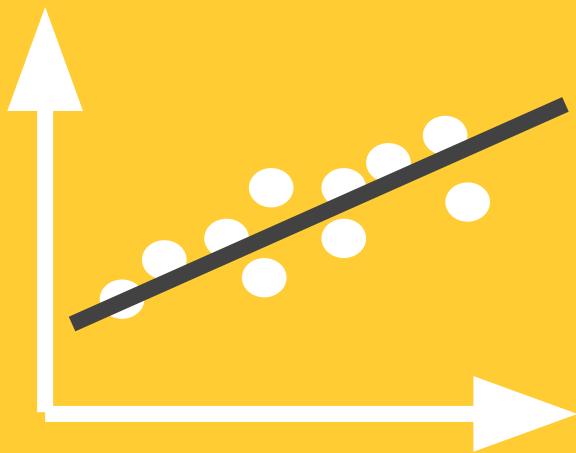




Regression Introduction

Exercise

- **What are the factors that determine whether a user would like a movie?**

- User location
- Movie duration
- Movie release date
- User gender
- Movie genre(s)
- User's rating
- User age group
- ...

movie id	movie title	release date	video release date	IMDb URL	unknown	Action	Adventure	Animation	Childrens	Comedy	Crime
1	Toy Story (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?Toy%20Story%20...	0	0	0	1	1	1	0
2	GoldenEye (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?GoldenEye%20(...	0	1	1	0	0	0	0
3	Four Rooms (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?Four%20Rooms%...	0	0	0	0	0	0	0
4	Get Shorty (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?Get%20Shorty%...	0	1	0	0	0	1	0
5	Copycat (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?Copycat%20(1995)	0	0	0	0	0	0	1

Exercise

- Can you predict ratings with these features?

Exercise

- What are simple ways to predict what a user would rate a movie?

Model 1 - Average

- **Use an average of the user's past ratings**

Model 1 - Average

- **What if our prediction is wrong?**
- **We could recommend the wrong movie**

Model 1 - Average

- How do we know if this model is accurate?
- Use an error metric to calculate error

Calculate Error

- **Simple method:**
- **Find difference between predicted and actual values**

Mean Squared Error

- Square errors so they don't cancel out
- Divide the result by the number of data points
 - So the error isn't dependent on number of data point

How do we improve the model?

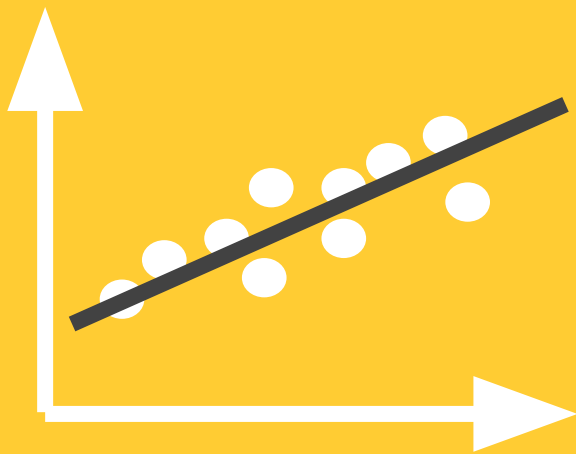
- Use a feature to group average ratings

Using a Feature - Example

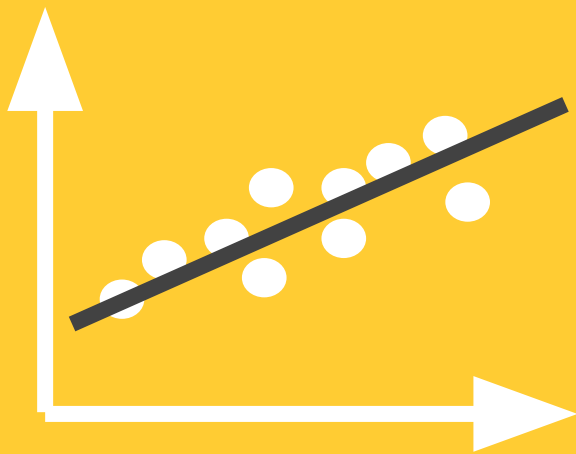
- Calculate the average of a movie per genre

Using a Feature - Example

- What if there are multiple features on which a rating depends?
- Use regression



**What is
Regression?**



**What is
Regression?**

Supervised Learning: Regression

“Supervised”?

- **A set of samples where the desired labels are known**

What is the main goal of supervised learning?

- **Learn a model from labeled training data that allows for predictions on new unseen data**

Supervised Learning: Regression

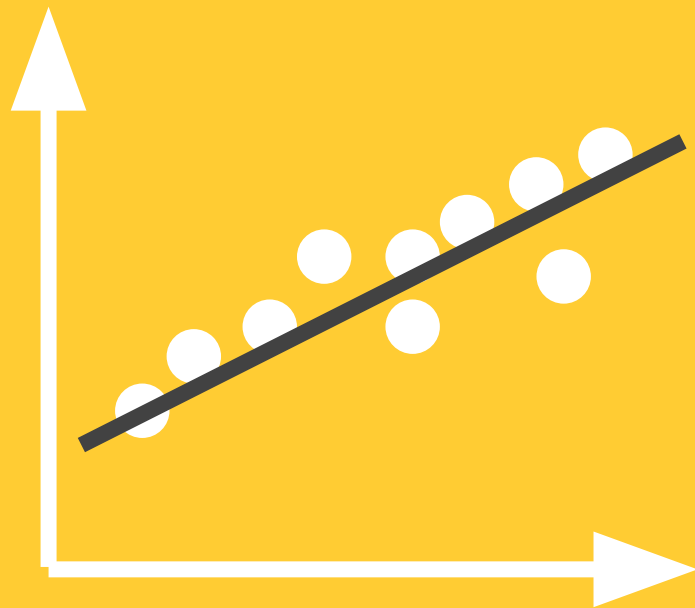
- **Given a training set of labeled input-output pairs**
 - $(x_1, y_1), (x_2, y_2) \dots (x_N, y_N)$
 - x and y can be any value
 - each y is generated by an unknown function $y = f(x)$

Supervised Learning: Regression

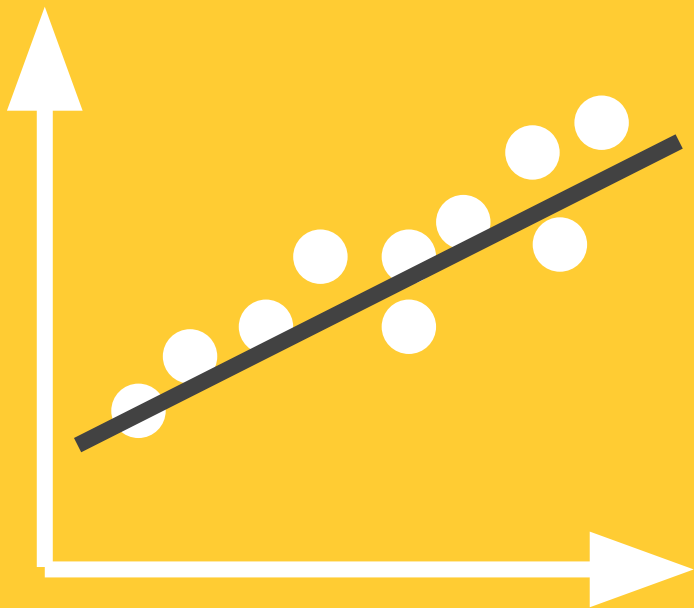
- What is the goal?
 - Find a hypothesis h that approximates the true function f

What is regression?

- Subcategory of supervised learning
- What is the goal?
 - predict continuous outcomes



Linear Regression



- Data generated by true function f + noise

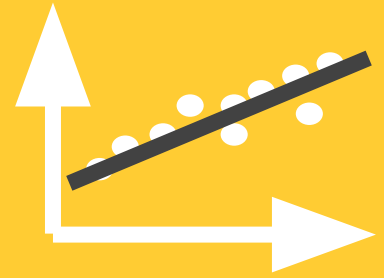


Hypothesis h

- Equation $y = mx + b$
- M = slope
- B = y-intercept

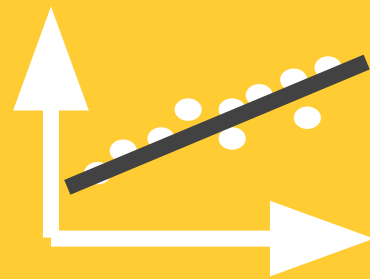
Regression

- Given a number of features and a continuous outcome, find a relationship between the features to predict an outcome

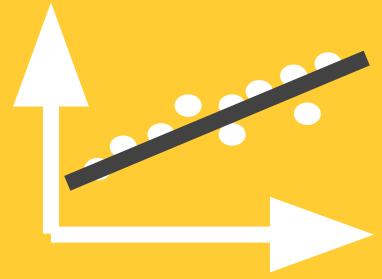


- Assumes the regression function is linear
 - Simple
-
- Often adequate and interpretable description of how inputs affect outputs

- **Sometimes outperform non linear models!**



- **Especially with:**
 - small training datasets, or
 - low signal to noise ratio, or
 - sparse data

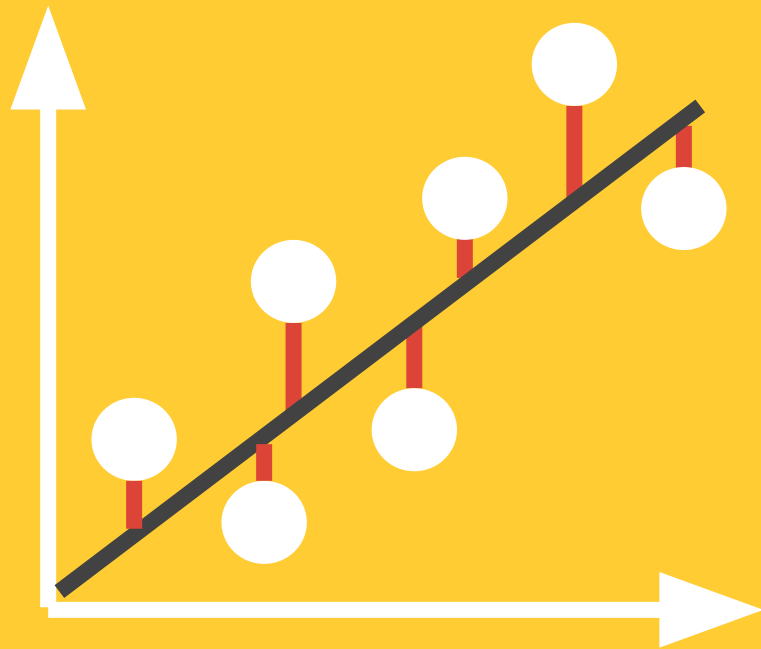


Linear Model Fit By Least Squares Method

- Prediction method
- Mainstay of statistics
- One of the most important tools

Linear Regression: Machine Learning

- What is the goal?
 - minimize **error**

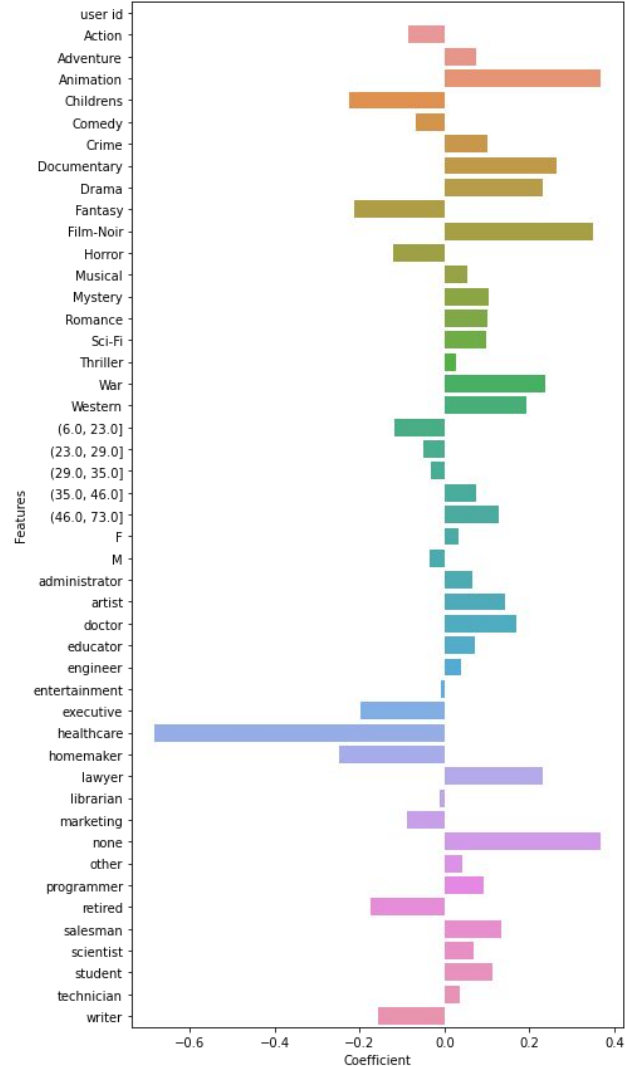


Every model is in danger of overfitting

- Overtrain
- As you add more model parameters, complexity increases, which overfits
- How do we find the sweet spot of training?

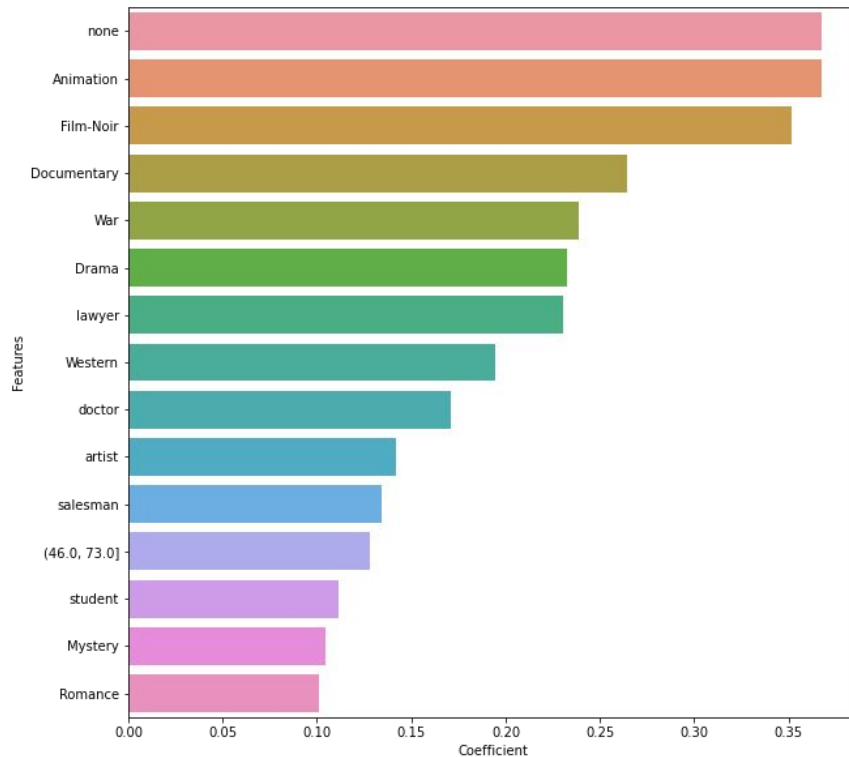
Ridge Regression

- Uses regularization
- Keep the same number of features but reduce magnitude of coefficients



Why do we want to reduce magnitude of coefficients?

- If one feature has a very high coefficient, the prediction is more driven by that feature

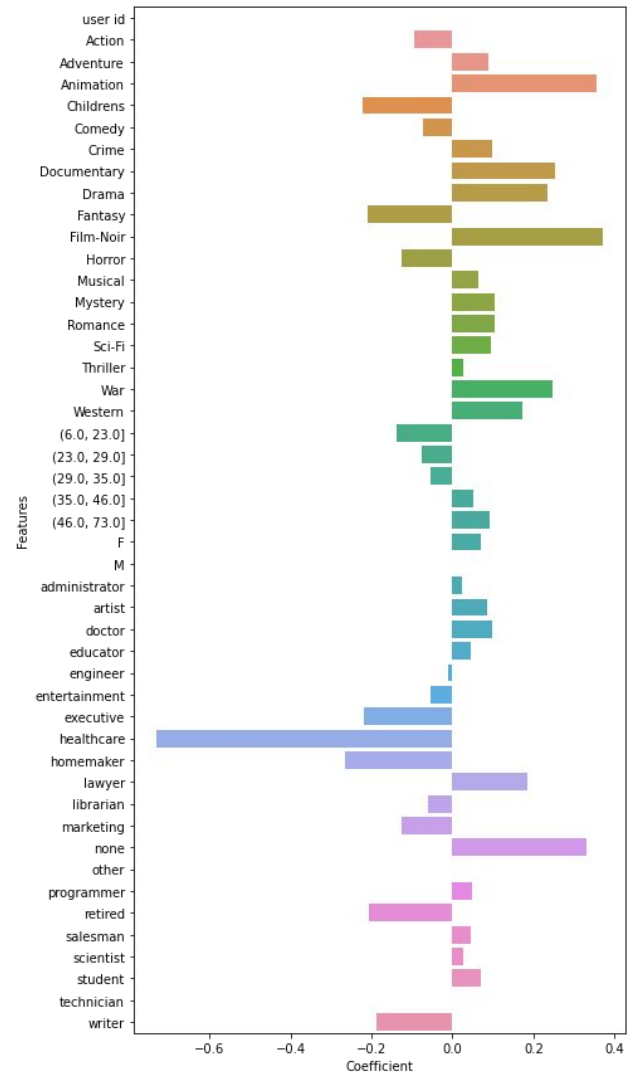


Lasso Regression

- **Similar to Ridge regression**
- **Selects only some features while reducing coefficients of others to zero**
 - Feature selection

Lasso Regression

- Typically used when you have more features



Example - Which one do we use?

- **Suppose we had a large dataset with 1,000 features**
 - Some independent features correlate with others

Example - Ridge

- **Model retains all features**
- **Shrinks coefficients**
- **Model will still remain complex because there are 1,000 features**
 - Can lead to poor model performance

Example - Lasso

- When we have correlated features, lasso retains only 1
- Sets the other correlated features to 0

Solutions

- **Test both models**
- **Combine both models**

What is Linear Regression?

